Appel à candidatures

CONTRATS DOCTORAUX EN INTELLIGENCE ARTIFICIELLE

Sujet de thèse doctorat 2023-2026

Injection de connaissances pour apprendre des représentations multimédia explicables

Mots clés : apprentissage profond, modèles explicables, injection de connaissances, données multimédia, confiance de l'intelligence artificielle.

1 Contexte et état de l'art

1.1 Contexte

L'utilisation croissante de modèles appris à partir de données dans des systèmes critiques ou sensibles accentue la nécessité de développer des méthodes permettant aux utilisateurs humain de mieux en comprendre le fonctionnement interne et les processus de décision [7, 4]. L'avènement de l'internet des objets (*Internet of Things : IoT*) [20] et le nombre croissant d'appareils connectés [1] ne fera qu'accentuer cette tendance.

Cette problématique est particulièrement saillante dans les domaines de la santé, de la défense et du transport (véhicules autonomes) où les décisions prises ont des conséquences directes sur la vie humaine. De manière plus générale, se reposer sur des algorithmes d'IA pour prendre des décisions critiques ne peut pas exempter de toute responsabilité légale et financière en cas d'erreur. Il est donc important d'être capable de comprendre l'origine de telles erreurs, de déterminer quelle part incombe à la conception ou à l'utilisation d'un algorithme [6, 4]. Plus en amont, la compréhension de ces failles permet aussi de s'en prémunir. Rendre les modèle neuronaux plus intuitifs, transparents et interprétables est également crucial quand ils sont utilisés massivement pour décider des informations les plus mises en avant sur les réseaux sociaux et dans les média, dans le domaine financier et économique (octroi de prêts, décision d'achat/vente, arbitrages), voire dans l'assistance au domaine judiciaire. Dans de tels contextes, identifier et corriger d'éventuels biais indésirables participe probablement à maintenir un aspect démocratique de sociétés faisant appel à ces algorithmes [17].

1.2 État de l'art

L'explicabilité des modèles appris fait l'objet de nombreux travaux scientifiques, dans plusieurs sous-champs de l'intelligence artificielle [11, 4, 18].

Intégration de connaissances dans des tâches multimédia. Plusieurs travaux ont porté sur des tâche nécessitant l''intégration des modalités visuelles et textuelles, comme légender automatiquement des images (et réciproquement, l'illustration visuelle d'un texte) ou les systèmes de question-réponse visuelle (*visual question answering* – VQA) nécessitant de répondre à une question (textuelle) portant sur une image. Les extensions multimodales des tâches traditionnelles de traitement du langage naturel (NLP), tel que la traduction automatique multimodale (MMT) [13],

1. Selon IoT Analytics, il y avait 14,4 milliards de connectés appareils en 2022

la reconnaissance multimodale d'entités nommées (MNER) [16, 24] et la reconnaissance d'entités multimodales (MEL) [1] bénéficient aussi de l'intégration vision-langage. Le principal défi pour ces tâches consistait initialement à aligner les représentations des deux modalités dans un espace commun implicite ou explicite [3, 21] mais d'autres travaux [23, 22, 15, 19] ont rapidement proposé une formulation de tâches VQA basée sur les connaissances et nécessitant des capacités de raisonnement.

Ces approches incluent des questions nécessitant des connaissances externes « de bon sens » pour être correctement répondues, car le contenu visuel n'intègre que des informations partielles. Des travaux plus récents ont proposé d'aller au delà du « sens commun », qui est vague, difficile à définir, et fluctuant selon les individus et communautés. Shah *et al*. a proposé de répondre à des des questions portant sur des entités nommées de type « personnes » à partir d'informations texte et image. Lerner *et al*. [12] a généralisé le problème en proposant la tâche KVQAE (Knowledge-based Visual Question Answering about named Entities), une formulation de problème VQA où répondre questions nécessite des connaissances sur les entités nommées définies dans une base de connaissances (KB). Ils ont proposé l'ensemble de données ViQuAE, qui couvre des centaines de types d'entités. L'accroissement récent des performances des grands modèles de langage stimule la poursuite des recherches sur cette tâche [5].

Une direction complémentaire poursuivie récemment pour le raisonnement multimodal texte et image est l'utilisation des réseaux modulaires (Neural Module Networks - NMN). Pour rendre le processus de raisonnement plus transparent et plus proche de l'interprétation humaine, les NMN compositionnels [9, 14] effectuent un raisonnement en décomposant une tâche de raisonnement complexe en plusieurs sous-tâches plus faciles, chacune implémentée comme un sous-réseau spécifique [2]. Cependant, les approches existantes ont certains inconvénients, qui suggèrent autant de pistes de developpement futur : pas de garantie sur la cohérence du raisonnement entre les différentes sous-modules (entre la question globale et ses sous-questions) [10], pas de robustesse au changement du contexte visuel [8] et pas de supervision mutuelle (la convergence du chaque module vers la fonction qui lui a été attribué n'est pas garantie – ce qui nuit à l'interprétabilité du système) [2].

2 Objectifs et approche

Dans le cadre de cette thèse, nous explorerons la capacité à accroître la confiance envers les modèles appris à partir de données visuelles et textuelles, en se reposant sur des connaissances externes structurées. Plus précisément, nous viserons à injecter des connaissances pour apprendre des modèles et des représentations explicables, avec une attention directe sur les données de type multimédia. Plusieurs directions de recherche sont envisageables, cf. l'état de l'art en haut.

Au delà du sujet lui-même, une thèse en IA est aujourd'hui en grande partie jugée par les publications scientifiques issues du travail doctoral. Aussi, un objectif pratique important est de publier le travail de thèse dans des conférences et journaux scientifiques du meilleur niveau dans les domaines de la vision par ordinateur (CVPR, ICCV, ECCV), du multimédia (ACM Multimédia), de l'apprentissage (ICLR, NeurIPS, ICML) ou du traitement du langage (ACL, EMNLP).

Ces publications scientifiques pourront résulter de développements théoriques et/ou expérimentaux

- Clarifier théoriquement les différentes dimensions de la confiance envers les modèles d'IA, en objectivant sa caractérisation en terme de transparence, explicabilité, intelligibilité, interprétabilité, robustesse, causalité, etc.
- Développer des méthodes permettant d'accroître la confiance des utilisateurs humains envers les modèles appris à partir de données visuelles et/ou textuelles, en se reposant sur des données structurées externes de référence.

— Valider expérimentalement les méthodes sur des corpus académiques de référence. Il sera possible d'éventuellement participer au développement de tels benchmarks si cela peut aboutir à une contribution scientifique significative dans un temps raisonnable. Il pourra aussi être envisagé d'évaluer les méthodes proposées sur des cas d'usages plus « réalistes », toujours si cela peut permettre une valorisation sous forme d'article scientifique de bon niveau.
— Développer des liens avec des champs de recherche connexes de la conception de modèles d'intelligence artificielles, ayant traits aux questions éthiques que ces modèles soulèvent. on pensera notamment au domaine juridiques et aux sciences humaines et sociales (SHS).

3 Planning prévisionnel

Le doctorat est un programme de trois ans au cours duquel le candidat sera à la fois inscrit en tant qu'étudiant et employé par le CNAM List à Paris pour un contrat à durée déterminée. La date de début est prévue pour l'automne 2023 avec une soutenance et une remise de diplôme à l'automne 2026. Le salaire net devrait se situer autour de 1800€/mois

Le plan habituel d'un doctorat est le suivant :

— 3-4 mois : recherche bibliographique afin de mieux comprendre le contexte, d'identifier les opportunités de recherche et de définir des orientations précises pour les années suivantes. Cette période est également utilisée pour commencer de "petites expériences" afin d'apprendre à manier les outils classiques de vision, apprentissage, NLP et d'approfondir la compréhension du candidat de l'état de l'art en reproduisant des contributions récentes.
— Première année : le reste de la première année de la thèse est consacré à l'exploration d'idées de recherche pour aboutir à un premier prototype d'une approche qui pourrait faire l'objet d'une publication dans une conférence internationale.
— Deuxième année : elle est consacrée au développement intense de la recherche, validée par des publications de niveau croissant, jusqu'aux meilleures conférences du domaine (CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML, IJCAI, ECAI, ACL, EMNLP, NAACL...).
— Troisième année : la dernière année du doctorat est divisée en deux. Le premier semestre est utilisé pour « finaliser » le travail de recherche, y compris les expériences les plus ambitieuses combinant plusieurs travaux développés au cours des années précédentes et publiés dans des conférences ou des revues de haut niveau. Le second semestre est consacré à la rédaction du manuscrit (avec une mise en page de la thèse proposée environ 6 mois avant la fin) et à l'insertion professionnelle (développement plus actif d'un réseau, début de la recherche d'emploi ou de post-doc, etc.)

4 Profil recherché

Candidat(e) titulaire d'un diplôme de master (ou équivalent) :

— Bonne maîtrise du domaine de l'apprentissage statistique et apprentissage profond
— Bonnes connaissances de Python et des librairie Deep Learning PyTorch et/ou Tensorflow
— La maîtrise de l'anglais technique est essentielle

Envoyez vos candidatures (avec CV, lettre de motivation et relevé de notes) à Marin Ferecatu (marin.ferecatu@cnam.fr) et Herve Le Borgne (herve.le-borgne@cea.fr).

5 Organisation

La thèse se déroulera en co-encadrement entre l'équipe de recherche Vertigo du laboratoire CEDRIC (CNAM, Paris) et le laboratoire AI, Vision et Langage du CEA-List.

Plusieurs enseignants-chercheurs et doctorants des équipes d'encadrement travaillent sur les applications des réseaux de neurones profonds pour la compréhension et la structuration des de données de type multimédia (image/vidéo/son).

6 Encadrement

Dr. Marin Ferecatu
Maître de conférences, HDR, CNAM Paris
2 rue Conté, 75003 Paris
Email : marin.ferecatu@cnam.fr

Dr. Hervé Le Borgne
CEA Saclay - Nano-INNOV
DRT/LIST/DIASI/SIALV Hervé Le Borgne – LASTI
Bat 861 - PC 184 - F91191 Gif-sur-Yvette Cedex France
Email : herve.le-borgne@cea.fr

7 Informations complémentaires

Type de contrat : thèse doctoral (36 mois)
Date de début : Octobre/Novembre 2023
Rémunération : approx. 1800€/mois
Niveau d'études : Bac+5 / Master

Références

[1] Omar Adjali, Romaric Besancon, Olivier Ferret, Herve Le-borgne, and Brigitte Grau. Multimodal entity linking for tweets. In *European Conference on Information Retrieval (ECIR)*. Springer, 2020.

[2] Wafa Aissa, Marin Ferecatu, and Michel Crucianu. Curriculum learning for compositional visual reasoning. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5 : VISAPP,*, pages 888–897. SciTePress, 2023.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa : Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58 :82–115, 2020.

[5] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions ?, 2023.

[6] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai) : A survey. *CoRR*, abs/2006.11371, 2020.

[7] Diogo Vaz de Carvalho, Eduardo Marques Pereira, and Jaime S. Cardoso. Machine learning interpretability : A survey on methods and metrics. *MDPI Electronics (section : Artificial Intelligence)*, 2019.

[8] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. Swapmix : Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5078–5088, June 2022.

[9] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason : End-to-end module networks for visual question answering. In *ICCV*, 2017.

[10] Chenchen Jing, Yunde Jia, Yuwei Wu, Xinyu Liu, and Qi Wu. Maintaining reasoning consistency in compositional visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5108, June 2022.

[11] Journée inter-GdR ISIS, IA, IGRV et club EEA. IA et réseaux de neurones profonds, ouvrir la boite noire : du modèle explicable à la synthèse et présentation d'explications en signal et image., March 2023.

[12] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120, 2022.

[13] Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 6327–6337, Dublin, Ireland, 2022. Association for Computational Linguistics.

[14] Guohao Li, Xin Wang, and Wenwu Zhu. Perceptual visual reasoning with knowledge propagation. In *ACM MM*, MM '19, page 530–538, New York, NY, USA, 2019. ACM.

[15] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA : A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019.

[16] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[17] AI4MEDIA project. Initial analysis of the legal and ethical framework of trusted AI, 2022.

[18] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai : A survey on interpreting the inner structures of deep neural networks, 2023.

[19] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa : A benchmark for visual question answering using world knowledge. *arXiv preprint arXiv :2206.01718*, abs/2206.01718, 2022.

[20] Wiebke Toussaint and Aaron Yi Ding. Machine learning systems for intelligent services in the iot : A survey. *CoRR*, abs/2006.04950, 2020.

[21] Thi Quynh Nhi Tran, Hervé Le Borgne, and Michel Crucianu. Aggregating image and text quantized correlated components. In *IEEE/CVF Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 6 2016.

[22] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa : Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10) :2413–2427, 2017.

[23] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1290–1296. ijcai.org, 2017.

[24] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352, Online, 2020. Association for Computational Linguistics.